Mathematical Toolkit **Spring 2023**

Lecture 14: May 8, 2023

Lecturer: Ali Vakilian (notes from Madhur Tulsiani)

1 Probability over (uncountably) infinite probability spaces

Extending the idea of defining a probability *for each outcome* becomes problematic when we try to extend it to uncountably infinite spaces. For example, let $\Omega = [0,1]$. Let $\nu : [0,1] \to [0,1]$ be a function, which we want to think of as a probability distribution. Define the set

$$S_n \;=\; \left\{ x \in [0,1] \mid \nu(x) \geq \frac{1}{n} \right\}.$$

Since we want the total probability to add up to 1, we must have $|S_n| \leq n$. Also,

$$\text{Supp}(\nu) \;=\; \{x \in [0,1] \mid \nu(x) > 0\} \;\subseteq\; \cup_{n=1}^{\infty} S_n \,.$$

Since $\cup_{n=1}^{\infty} S_n$ is a countable set, $\nu(x) > 0$ only for countably many points x. Hence, it is problematic to think of the probability of the outcome x, for each $x \in [0,1]$. This can be resolved by only talking of probabilities of *events* for an allowed set of events obeying some nice properties. Such a set is known as a σ-algebra or a σ-field.

Definition 1.1 *Let 2^{Ω} denote the set of all subsets of Ω. A set $\mathcal{F} \subseteq 2^{\Omega}$ is called a* σ-field *(or σ-algebra) if*

1. $\emptyset \in \mathcal{F}$.
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ *(where $A^c = \Omega \setminus A$).*
3. *For a (countable) sequence $A_1, A_2, \ldots$ such that each $A_i \in \mathcal{F}$, we have $\cup_i A_i \in \mathcal{F}$.*

We then think of the sets in $\mathcal{F}$ as the allowed events. We can now define probabilities as follows.

Definition 1.2 *Given a σ-field $\mathcal{F} \subseteq 2^{\Omega}$, a function $\nu : \mathcal{F} \to [0,1]$ is known as a* probability measure *if*

1. $\nu(\emptyset) = 0$.

2. $\nu(E^c) = 1 - \nu(E)$ *for all* $E \in \mathcal{F}$.

3. *For a (countable) sequence of disjoint sets* $E_1, E_2, \ldots$ *such that all* $E_i \in \mathcal{F}$*, we have*

$$\nu\left(\cup_i E_i\right) \;=\; \sum_i \nu(E_i)\,.$$

Note that the above definition do not say anything about unions of an uncountably infinite collection of sets. We can of course define probability measures on $\mathcal{F} = 2^\Omega$ and hence define $\nu(x)$ for all $x \in \Omega$. However, as we saw above, such measures will only have $\nu(x) > 0$ countably many x. Consider the following example.

Example 1.3 *Let* $\Omega = [0,1]$ *and* $\mathcal{F} = 2^\Omega$*. Let* $T = \left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\}$*. For each* $S \in \mathcal{F}$*, define*

$$\nu(S) \;=\; \frac{|S \cap T|}{4}\,.$$

In the above example, $\nu(x) > 0$ only for the points in a finite set T, which is very restrictive. We would like to formalize intuitive notions like the "uniform distribution" on the space $\Omega = [0,1]$: a probability measure that satisfies $\nu([a,b]) = b - a$ for $a, b \in [0,1]$ or more generally, for any event E and a circular shift $E \oplus x$ for $x \in [0,1]$, we want $\nu(E) = \nu(E + x)$. It is a non-trivial result that such a probability measure indeed exists. This probability measure is known as the *Lebesgue* measure and is defined over a σ-algebra known as the *Borel* σ-algebra. The Borel σ-algebra does not contain all subsets of $[0,1]$ but does contain all intervals $[a,b]$. In fact, one can use the axiom of choice to show that we *cannot* include all subsets. The reason is that countable unions of very "thin" disjoint sets can reconstruct a "thick" set.

Proposition 1.4 *Let* $\Omega = [0,1]$*. A measure satisfying the requirement that* $\nu(E) = \nu(E + x)$ *for all* $E \in \mathcal{F}$ *cannot be defined over the* σ*-algebra* $\mathcal{F} = 2^\Omega$*.*

Proof: For the sake of contradiction, assume that such a measure exists. Let $\mathcal{B}$ be the set of numbers in $[0,1)$ with a finite binary expansion, and define the equivalence relation between points $x, y \in [0,1]$:

$$x \sim y \qquad \text{if } \exists b \in \mathcal{B} \text{ such that } x = y \oplus b.$$

Thus x and y are equivalent if we can change only finitely many of the binary expansion of one, to get the other. Let $[x]$ denote one such equivalence class. Note that since there are countably many elements in $\mathcal{B}$, $[x]$ is also countable. In particular, $[0] = \mathcal{B}$. Because an equivalence defines a partition, it follows that there must be uncountably many distinct $[x]$'s that are furthermore disjoint. Now, by the axiom of choice, construct a set V that selects only one element from each such distinct $[x]$. V thus has uncountably many

elements, but in some sense, is "thin". Consider all the circular shifts of V of the form $V \oplus b$ for $b \in \mathcal{B}$. These are disjoint, since we never recreate the same element within the equivalence class of a given point x (why?) nor jump from the equivalence class of x to that of another. Furthermore as b varies, each x recreates its entire equivalence class, and it follows that:

$$\bigcup_{b \in \mathcal{B}} V \oplus b = [0, 1].$$

So now we ask, what can $\nu(V)$ be? It certainly cannot be positive, since otherwise $\nu([0,1]) = \sum_{b \in \mathcal{B}} \nu(V \oplus b) = \sum_b \nu(V) = \infty$. But it cannot be zero either, since otherwise $\mathbb{P}([0,1]) = \sum_b \nu(V) = 0$. This is a contradiction. ■

What went wrong? This is a very involved debate, but essentially the issue is an interaction between countable additivity and our ability to have created V in the first place. The attitude of probability theory can be interpreted as either denying that such sets exists, or accepting that they do exist, but refusing to define the probability measure over them. The latter turns out to be much more productive, because the notion of restricting the probability measure to only given subsets has many versatile uses, including a generalization of the notion of conditioning.

1.1 Random Variables

Recall that to define a random variable, we need to define a σ-algebra on the range of the random variable. A random variables is then defined as a *measurable* function from the probability space to the range: functions where the pre-image of every subset in the range σ-algebra is an event in $\mathcal{F}$.

An important case is when the range is $[0, 1]$ or $\mathbb{R}$. In this case we say that we have a *real-valued* random variable, and we use the Borel σ-algebra unless otherwise noted. For countable probability spaces, we wrote the expectation of a real-valued random variable as a sum. For uncountable spaces, the expectation is an integral with respect to the measure.

$$\mathbb{E}[X] \;=\; \int_\Omega X(\omega) d\nu \,.$$

The definition of the integral with respect to a measure requires some amount of care, though we will not be able to discuss this in much detail. Let ν be any probability measure over the space $\mathbb{R}$ equipped with the Borel σ-algebra. Define the function F as

$$F(x) \;:=\; \nu((-\infty, x]) \,,$$

which is well defined since the interval $(-\infty, x]$ is in the Borel σ-algebra. This can be used to define a random variable X such that $\mathbb{P}[X \leq x] = F(x)$. The function F is known as the cumulative distribution function (CDF) of X.

When the function F has the form

$$F(x) \;=\; \int_{-\infty}^{x} f(z)dz\,,$$

then f is called the density function of the random variable X. In this case, one typically refers to X as a "continuous" random variable. To calculate the above expectation for continuous random variables, we can use usual (Lebesgue) integration:

$$\mathbb{E}\,[X] \;=\; \int_{\mathbb{R}} xf(x)dx\,.$$

(The notion of density can be extended to between any two measures, via the Radon-Nikodym theorem. In that context, the density f of a continuous random variable is referred to as the Radon-Nikodym derivative with respect to the Lebesgue measure. In the earlier example with the measure concentrated on the finite set T, the probability of each point is the Radon-Nikodym derivative with respect to the counting measure of T: $\nu_T = \sum_{t\in T} \delta_t$.)

2 Gaussian Random Variables

A Gaussian random variable X is defined through the density function

$$\gamma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,,$$

where μ is its mean and σ^2 is its variance, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$. To see the definition gives a valid probability distribution, we need to show $\int_{-\infty}^{\infty} \gamma(x)dx = 1$. It suffices to show for the case that $\mu = 0$ and $\sigma^2 = 1$. First we show the integral is bounded.

Claim 2.1 *$I = \int_{-\infty}^{\infty} e^{-x^2/2}dx$ is bounded.*

Proof: We see that

$$I = \int_{-\infty}^{\infty} e^{-x^2/2}dx = 2\int_{0}^{\infty} e^{-x^2/2}dx \le 2\int_{0}^{2} 1dx + 2\int_{2}^{\infty} e^{-x}dx = 4 + 2e^{-2}\,,$$

where we use the fact that I is even and after $x = 2$, $e^{-x^2/2}$ is upper bounded by e^{-x}. ∎

Next we show that the normalization factor is $\sqrt{2\pi}$.

Claim 2.2 *$I^2 = 2\pi$.*

Proof:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2/2} dx \int_{-\infty}^{\infty} e^{-y^2/2} dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy \\ &= \int_0^{\infty} \int_0^{2\pi} e^{-r^2/2} r dr d\theta \quad (\text{let } x = r\cos\theta \text{ and } y = r\sin\theta) \\ &= 2\pi \int_0^{\infty} e^{-s} ds \quad (\text{let } s = r^2/2) \\ &= 2\pi \,. \end{aligned}$$

■

This completes the proof that the definition gives a valid probability distribution. We prove a useful lemma for later use.

Lemma 2.3 *For $X \sim \mathcal{N}(0,1)$ and $\lambda \in (0, 1/2)$,*

$$\mathbb{E}\left[e^{\lambda \cdot X^2}\right] = \frac{1}{\sqrt{1-2\lambda}} \,.$$

Proof:

$$\begin{aligned} \mathbb{E}\left[e^{\lambda \cdot X^2}\right] &= \int_{-\infty}^{\infty} e^{\lambda \cdot x^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(1-2\lambda)x^2/2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \frac{dy}{\sqrt{1-2\lambda}} \quad (\text{let } y = \sqrt{1-2\lambda}x) \\ &= \frac{1}{\sqrt{1-2\lambda}} \end{aligned}$$

■

3 Johnson–Lindenstrauss Lemma

We will use concentration bounds on Gaussian random variables to prove the following important lemma.

Lemma 3.1 (Johnson–Lindenstrauss) *Let $\mathcal{P}$ be a set of n points in $\mathbb{R}^d$. Let $0 < \varepsilon < 1$. For $k = \frac{8 \ln n}{\varepsilon^2/2 - \varepsilon^3/2}$, there exists a mapping $\varphi : \mathcal{P} \to \mathbb{R}^k$ such that for all $u, v \in \mathcal{P}$*

$$(1-\varepsilon)\|u-v\|^2 \leq \|\varphi(u) - \varphi(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2 \,.$$

The above lemma is useful for dimensionality reduction, especially when a problem has an exponential dependence on the number of dimensions. We construct the mapping φ as follows. First choose a matrix $G \in \mathbb{R}^{k \times d}$ such that each $G_{ij} \sim \mathcal{N}(0,1)$ is independent. Define

$$\varphi(u) = \frac{Gu}{\sqrt{k}}.$$

Note that by the above construction φ is oblivious, meaning that it doesn't depend on the points in $\mathcal{P}$, and it is linear. Before we prove the lemma, we will use the following fact several times.

Fact 3.2 *Let $Z = c_1 X_1 + c_2 X_2$, where $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim \mathcal{N}(0,1)$ are independent. Then $Z \sim \mathcal{N}(0, c_1^2 + c_2^2)$.*

The strategy of proving the lemma is to first prove that with high probability the lemma holds for any fixed two points and then apply union bounds to get the result for all pairs of points.

Claim 3.3 *Fix $u, v \in \mathcal{P}$. Let $w = u - v$. With probability greater than $1 - 1/n^3$, the following inequality holds,*

$$(1-\varepsilon) \cdot \|w\|^2 \;\le\; \|\varphi(w)\|^2 \;\le\; (1+\varepsilon) \cdot \|w\|^2 .$$

Proof: Recall that $\varphi(u) = \frac{Gu}{\sqrt{k}}$. Let

$$Z = \frac{k\|\varphi(w)\|^2}{\|w\|^2} = \frac{\sum_{i=1}^{k} (Gw)_i^2}{\|w\|^2}.$$

We need to show $(1-\varepsilon)k \le Z \le (1+\varepsilon)k$. We know that the sum of Gaussian random variables is still a Gaussian random variable, so $(Gw)_i = G_i w = \sum_{j=1}^{n} G_{ij} w_j$ is a Gaussian variable. Besides, $\mathsf{Var}\left[\sum_{j=1}^{n} G_{ij} w_j\right] = \sum_j w_j^2 = \|w\|^2$ according to Fact 3.2. In other words, $G_i w \sim \mathcal{N}(0, \|w\|^2)$. As a result, $Z = \sum_{i=1}^{k} \frac{(Gw)_i^2}{\|w\|^2} = \sum_{i=1}^{k} X_i^2$, where $X_i \sim \mathcal{N}(0,1)$. The expectation of each individual element in Gw is

$$\mathbb{E}\left[(Gw)_i^2\right] = \mathbb{E}\left[(G_i w)^2\right] = \mathbb{E}\left[\left(\sum_{j=1}^{n} G_{ij} w_j\right)^2\right] = \mathsf{Var}\left[\sum_{j=1}^{n} G_{ij} w_j\right] = \|w\|^2 .$$

In addition,

$$\mathbb{E}[Z] = \frac{\sum_{j=1}^{k} \mathbb{E}\left[(Gw)_i^2\right]}{\|w\|^2} = k .$$

Now we prove the concentration bound for Z. The proof is almost identical to Chernoff bound.

$$\begin{aligned}
\mathbb{P}\left[Z \geq (1+\varepsilon)k\right] &\leq \mathbb{P}\left[e^{\lambda Z} \geq e^{\lambda\cdot(1+\varepsilon)k}\right] \\
&\leq \frac{\mathbb{E}\left[e^{\lambda\cdot Z}\right]}{e^{\lambda\cdot(1+\varepsilon)k}} && \text{(by Markov's inequality)} \\
&= \frac{\mathbb{E}\left[e^{\lambda\cdot\sum_{i=1}^{k} X_i^2}\right]}{e^{\lambda\cdot(1+\varepsilon)k}} = \frac{\prod_{i=1}^{k}\mathbb{E}\left[e^{\lambda\cdot X_i^2}\right]}{e^{\lambda\cdot(1+\varepsilon)k}} && \text{(by the independence of } X_1,\ldots,X_k) \\
&= \frac{\prod_{i=1}^{k}\frac{1}{\sqrt{1-2\lambda}}}{e^{\lambda\cdot(1+\varepsilon)k}} && \text{(by Lemma 2.3)} \\
&\leq \left(\frac{e^{-2(1+\varepsilon)\lambda}}{1-2\lambda}\right)^{k/2} && \text{(assume } \lambda < 1/2) \\
&\leq \left(e^{-\varepsilon}(1+\varepsilon)\right)^{k/2} && \text{(let } \lambda = \frac{\varepsilon}{2(1+\varepsilon)}) \\
&\leq \left((1-\varepsilon+\frac{\varepsilon^2}{2})(1+\varepsilon)\right)^{k/2} && \text{(by Taylor expansion of } e^{-x}) \\
&\leq e^{-\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{2}\right)\frac{k}{2}} && \text{(by } 1+x \leq e^x)
\end{aligned}$$

We can derive the other side of the inequality in an analogous way. Thus, we have

$$\mathbb{P}\left[|Z-k| \geq \varepsilon k\right] \leq 2\cdot\exp\left(-\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{2}\right)\frac{k}{2}\right) \leq 2\cdot\exp\left(-3\ln n\right) = \frac{2}{n^3},$$

where we choose

$$k = \left\lceil \frac{6\ln n}{\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{2}} \right\rceil.$$

■

To prove Johnson–Lindenstrauss Lemma, we apply the union bound and get the desired result

$$\begin{aligned}
\mathbb{P}\left[\forall u,v\in\mathcal{P}, (1-\varepsilon)\|u-v\|^2 \leq \|\varphi(u)-\varphi(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2\right] &\geq 1-\binom{n}{2}\frac{2}{n^3} \\
&\geq 1-\frac{1}{n}.
\end{aligned}$$